Estimating Joint Distribution of Output and PC

jackmo

June 2023

1 Kernel Density Estimator

- Draw n samples of k from p(k)
- For a fixed α and k, we can calculate output, O, and 'PC' (direction describing most variance) α_0 where $\alpha_0 = \alpha^T k$
- We can draw n 2D gaussians centered around (O, α_0) with covariance matrix: ϵI (tune ϵ as hyperparameter). taking the sum of gaussians normalizing by 1/N, we have:

$$P(O, \alpha_0) = \frac{1}{N} \sum_i N((O, \alpha_0); (O_i, \alpha_i), \epsilon I) \tag{1}$$

$$P(\alpha_0) = \frac{1}{N} \sum_i N(\alpha_0; \alpha_i, \epsilon) \tag{2}$$

- Sample many α_0 according to $P(\alpha_0)$. At each α_0, we have a conditional distribution given by:

$$P(O|\alpha_0) = \frac{P(O, \alpha_0)}{P(\alpha_0)} = \frac{\sum_i N((O, \alpha_0); (O_i, \alpha_i), \epsilon I)}{\sum_i N(\alpha_0; \alpha_i, \epsilon)} \tag{3}$$

- Find the expectation $E(O|\alpha_0)$ for all n alpha given by (4). Find the variance of these expectations, $V(E(O|\alpha_0))$ and normalize by $V(O)$

$$E(O|\alpha_0) = \frac{\sum_i O_i N(\alpha_0; \alpha_i, \epsilon)}{\sum_i N(\alpha_0; \alpha_i, \epsilon)} \tag{4}$$

2 Coding

- Given: kdraws (numSamples x numK) drawn from p(k), alpha (numK x 1)
- calculate output and np.matmul(kdraws, alpha), denote these O,a_0

- Fix eps to some value. Determine how many samples we want from p(alpha), numAlpha
- for i in range(numAlpha), repeat the following steps:
- randomly choose an element of a_0. draw a point from a gaussian $N(a_0, eps)$, denote this value as a.
- loop through every element of a_0, $a_0[i]$. for each element, calculate the pdf: N(a;$a_0[i]$, eps). Save all these entries as an array, N_i
- Calculate np.dot(O, N_i)/np.sum(N_i) to get E(O—a_0)
- go up to step 3. Should have a matrix of expectations (numAlpha x numOutput), E. Take variance of this array over numAlpha to find the estimated SI for each output. Average over numOutput to find average SI. Normalize SI by variance

3 Gaussian Mixture Model

- We assume a gaussian mixture model has already been fitted to the data with n components. The probability space is 2-D with variables given by (O, α).
- Given weights, w, mean, μ, and covariance, Σ, the joint probability of the GMM is given by:

$$p(O, \alpha) = \sum_{k=1}^{n} w_k N(O, a; \mu_k, \Sigma_k)$$

(5)

Where μ_k and Σ_k are given by:

$$\mu_k = (\mu_O, \mu_\alpha) \tag{6}$$

$$\Sigma_k = \begin{bmatrix} \Sigma_{\alpha,\alpha} & \Sigma_{\alpha,O} \\ \Sigma_{O,\alpha} & \Sigma_{O,O} \end{bmatrix}$$

- The marginal probability of α is given by:

$$P(\alpha) = \sum_{k=1}^{n} w_k N(\alpha; \mu_{k,\alpha} \Sigma_{k,\alpha,\alpha}) \tag{8}$$

- The conditional probabilitiy is given by:

$$P(O|\alpha) = \frac{P(O,\alpha)}{P(\alpha)} = \frac{\sum_{k=1}^{n} w_k N(O,a;\mu_k,\Sigma_k)}{\sum_{k=1}^{n} w_k N(\alpha;\mu_{k,\alpha}\Sigma_{k,\alpha,\alpha})} \tag{9}$$

The Normal distribution in the numerator can be further simplified given that for a fixed k:

$$N(O,a;\mu,\Sigma) = N(\alpha;\mu_\alpha,\Sigma_{\alpha,\alpha})N(O;\mu_{O|\alpha},\Sigma_{O|\alpha}) \tag{10}$$

$$P(O|\alpha) = \sum_{k=1}^{n} \frac{w_k N(\alpha;\mu_\alpha,\Sigma_{\alpha,\alpha})}{\sum_{k=1}^{n} w_k N(\alpha;\mu_{k,\alpha}\Sigma_{k,\alpha,\alpha})} N(O;\mu_{O|\alpha},\Sigma_{O|\alpha}) \tag{11}$$

The expectation of this distribution is given by the following:

$$E(O|\alpha) = \sum_{k=1}^{n} \frac{w_k N(\alpha;\mu_\alpha,\Sigma_{\alpha,\alpha})}{\sum_{k=1}^{n} w_k N(\alpha;\mu_{k,\alpha}\Sigma_{k,\alpha,\alpha})} \mu_{O|\alpha} \tag{12}$$

$$\mu_{O|\alpha} = \mu_O + \frac{\Sigma_{O,\alpha}}{\Sigma_{\alpha,\alpha}}(\alpha - \mu_\alpha) \tag{13}$$

Now we can draw α n times, calculate Eqtn 12 for a given α, and then find the variance of the expectations. The estimated sensitivity index is given by:

$$SI = \frac{Var(E(O|\alpha))}{Var(O)} \tag{14}$$

$$\text{Var(O)} = \sum_{k=1}^{N} w_k \cdot (\Sigma_{O,Ok} + \mu_{O,k}^2) - \left(\sum_{i=k}^{N} w_k \cdot \mu_k\right)^2 \tag{15}$$

4 GMM Higher Dimensions

- If we have multiple 'principle components' the formulae in the previous section can be easily adjusted
- For N PC's, we can treat α as an N-dimensional vector

$$E(O|\alpha) = \sum_{k=1}^{n} \frac{w_k N(\alpha;\mu_\alpha,\Sigma)}{\sum_{k=1}^{n} w_k N(\alpha;\mu_{k,\alpha}\Sigma_{k,\alpha,\alpha})} \mu_{O|\alpha} \tag{16}$$

5 GMM Algorithm

- Fit a GMM with nComponents to the data, yielding weights, covariances, and means for the gaussian mixture
- Draw nSamples number of samples from the GMM.
- For each drawn sample, calculate $E(O|\alpha)$ given in Eqtn 12.
- Calculate the variance of all the expectations.
- Calculate the total variance of O using Eqtn 15 and normalize the variance $Var(E(O|\alpha))$ by the total variance